Plant yield prediction in indoor farming using machine learning

Anjali Ashok
Mary Adesoba

Supervisor: Jonas Karlsson
Examiner: Juhee Bae

Master Degree Project (15 ECTS) in Informatics with a specialization in Data Science/Privacy, Information and Cyber Security
Spring term 2023
Abstract

Agricultural industry has started to rely more on data driven approaches to improve productivity and utilize their resources effectively. This thesis project was carried out in collaboration with Ljusgårda AB, it explores plant yield prediction using machine learning models and hyperparameter tweaking. This thesis work is based on data gathered from the company and the plant yield prediction is carried out on two scenarios whereby each scenario is focused on a different time frame of the growth stage. The first scenario predicts yield from day 8 to day 22 of DAT (Day After Transplant), while the second scenario predicts yield from day 1 to day 22 of DAT and three machine learning algorithms Support Vector Regression (SVR), Long Short Time Memory (LSTM) and Artificial Neural Network (ANN) were investigated.

Machine learning model’s performances were evaluated using the metrics; Mean Square Error (MSE), Mean Absolute Error (MAE), and r-squared. The evaluation results showed that ANN performed best on MSE and r-squared with dataset 1, while SVR performed best on MAE with dataset 2. Thus, both ANN and SVR meets the objective of this thesis work. The hyperparameter tweaking experiment of the three models further demonstrated the significance of hyperparameter tuning in improving the models and making them more suitable to the available data.

Keywords: Yield prediction, Machine Learning, Hyperparameter tweaking, Support Vector Regression, Long Short-Term Memory, Artificial Neural Network
# Table of Contents

1 Introduction  
1.1 Problem definition 2  
1.2 Aim and objectives 3  

2 Background  
2.1 Preliminaries  
   2.1.1 Support vector regression (SVR) 4  
   2.1.2 Long Short-Time Memory (LSTM) 5  
   2.1.3 Artificial neural network (ANN) 6  
   2.1.4 Performance metrics 6  
2.2 Research background 7  

3 Method  
4 Implementation of experiment  
4.1 Dataset 11  
4.2 Implementation of SVR (RQ1) 12  
4.3 Implementation of LSTM (RQ2) 13  
4.4 Implementation of ANN (RQ2) 14  
4.5 Statistical Analysis (RQ3) 14  

5 Results of experiment  
5.1 Results of support vector regression (RQ1) 14  
5.2 Results of long short-time memory (RQ2) 15  
5.3 Results of artificial neural network (RQ2) 15  
5.4 Results of statistical analysis (RQ3) 16  
5.5 Analysis of the joint results 16  

6 Discussion  
6.1 Comparison to Previous Research 19  
6.2 Methods, implementation, and results 20  
6.3 Ethical and societal aspects 22  

7 Conclusion  
7.1 Future work 23  

References 24  

Appendices  
Appendix A 28  
Appendix B 30
1 Introduction

In the agriculture sector, plant growth is very important to consider. Traditional farming practices rely on soil type, application of fertilisers, weather conditions etc. The plants grow inside the soil, where the soil feeds the plants with certain nutrients like nitrogen, phosphorus, potassium and so on. Temperature, humidity and light also contribute to the growth of plants. However, some of the main challenges in the system of traditional farming are the inability to meet up with increasing food demand due to population growth in recent times, effect of climate on the crops, pest and disease in the farm and poor management of resources by the farmers [35]. These challenges have raised lots of concern globally in the health of people and has affected the farmers financially [40]. To improve the quality of yield in plants and increase the production of food in order to meet food demand, there have been several technological approaches such as aquaponics, hydroponics, vertical farming, aeroponics and drip irrigation are used in controlled environments. Indoor farming is a modern farming practice where plants grow within a controlled environment [14]. In recent times, controlled farming is gradually emerging and have gained a lot of popularity in being able to manage resources and produce quality plant yield [35].

Presently, the world population is recorded to be around 7.7 billion and it’s progressing gradually and estimated to rise to 8.5 billion by 2030 [40]. With this current population growth rate, food demand is expected to increase. Therefore, it is important to look into these issues on how to increase food production in relation to progressive rise in the world population by investigating the factors that affect plant growth, gaining more insights into the plant growth issues, understanding when to plan and being able to control and manage resources accordingly which will lead to increase in crop productivity and profit for agricultural companies. After business objectives are well defined, suitable machine learning tools can be employed to understand, analyse and develop the data generated by sensors with the use of accurate forecasting models. In recent times, there is gradual improvement in food production with various technologies being utilised. To achieve and sustain progressive agriculture, research has been carried out in vertical farming and hydroponics. Vertical farming is a system whereby plants are grown under a well-monitored or controlled environment to ensure optimal plant growth [6]. Hydroponics method uses soilless agriculture where plants are grown in water that contains mineral nutrient [27]. According to [18], they discussed that the growth of plants in a controlled system has been improved by hydroponics operations. A better output is achieved in the quality and quantity of crops using the hydroponics system [31]. However, there are challenges in implementing the technology that enables these farming systems, which includes the cost to set up the environment and the sensors at the early stage, also it takes a lot of time to find suitable and accurate proportions of resources that is required or needed by the plants in the growing stage under the controlled environment.

Implementation of artificial learning has helped in achieving good productivity and factors that increase the vertical farming efficiency [7]. In this thesis, the objective is to predict plant yield by implementing three machine learning models which are SVR (Support Vector Regression), LSTM (Long Short-Term Memory) and ANN (Artificial Neural Network) using the data provided by the Ljusgårda AB. The models will be explored further through parameter tweaking to find its impact on the performance of the models. At the end of this experiment, the result would be helpful to the farmers, and this will support them in optimising their resources and increase the yield through machine learning.
1.1 Problem definition

The company Ljusgårda AB, the client of this thesis project, requests the prediction of their yield of lettuce plant variety “exanimo” for the days between 8 to 23 of the DAT stage (Day After Transplant) using machine learning algorithms. They currently have an automated system in place to track yield and climate variables. Through experimenting, the organization is also applying different nutrient solutions to improve yield. With this thesis, they will be able estimate yield based on various climate conditions and nutrients solutions using machine learning algorithms. This experiment also aimed to investigate the effect of tuning hyper parameters such as number of nodes in neural network and the choice of kernels in the SVM on the performance of the models.

In the company’s farm, from the germination stage to harvest stage of the lettuce plant (Fig 1), it takes a period of 35 to 40 days. The seed takes 1 day for the germination, and this is known as the germination stage. The next stage is the propagation stage, and it takes 15 to 20 days and this known as day after seedling stage (DAS), the last stage is the day after transplant (DAT) and this takes 17 to 22 days. The harvest is done between day 17 and 22 at the end of the DAT stage of growth period is from day 1 to 23. Ljusgårda AB is using hydroponics along with dripping irrigation technologies in their indoor farming setup. The drippers irrigate the lettuce plants with a set of volumes of water along with nutrients over a timeframe. These drippers are connected to rolling wagons which transport the lettuce plants from growing stage up to harvest stage. In the DAT period the sapling is transported to growing rooms and their growth is supported by controlled irrigation. Pipes carry water from one substrate to the next in order to soak the substrate. Thereafter, excess water goes down to tanks through the floor, which is recirculated and re-used. New nutrients are added to the water to balance the previously used nutrient, and the process is repeated all over again. Then this process is constantly monitored by an automated system to balance the nutrients. Ljusgårda AB uses sensors in the growing room for collecting various data over the plant growth stages. Light recipes are also used to control the light received by the plant in the growing rooms. These light recipes include green, blue, red, and far-red light at different intensity over a period of time.

Fig 1. Different stages of lettuce plant. (a) Germination stage. (b) Propagation stage. (c) plant substrate. (d, e) pipes carrying plant substrate. (f) lettuce in grow room (g) wagons during transportation. (h) harvesting of the plant

1 https://www.supernormalgreens.se/
1.2    Aim and objectives

The research objectives for this study are as follows:

1. How does the choice of the kernel in Support Vector Regression affect the performance of the model in predicting the yield?
2. In the neural networks LSTM and ANN, how does the number of neurons in hidden layer affect the model’s performance?
3. Based on the data, which one of the implemented models has a better performance?

The goal of RQ1 was to investigate the impact of kernels selection in the performance of SVR. SVR with different kernels were be implemented to see how the model performance is changing.

RQ2 seeks to explore the effect of parameters like number of layers and neurons in LSTM and ANN models. For this, LSTM and ANN model were experimented with different number of neurons in the hidden layer.

The objective of RQ3 was to find the model with best performance among the implemented models. Statistical analysis using kfold cross validation was used to confirm the stability of the models across different folds of data. Based on the results the model with tight confidence interval, and least error was chosen as the best model.

This study comprises various sections in order to explain the problem, implementation and evaluation of machine learning algorithms in this experiment. Section 1 gives an overview of the area of our thesis, along with the aims and objectives. Section 2 gives previous research that was carried out on this area and describes the algorithms used in this experiment. The approach that was followed in this thesis and its implementation is detailed in section 4. The results and discussion of the experiment was described in section 5 and 6 respectively. The last section is the conclusion and future work, and thereafter followed by references and appendices. Appendix A have information about the data provided by the company and the images of exploratory data analysis (EDA). Appendix B contains the contribution of individual group members in thesis writing and project.

2    Background

This background gives a brief description on the theory of machine learning models SVR, LSTM and ANN used in this experiment of plant growth prediction along with their performance metrics as well as previous related research background.

2.1    Preliminaries

Machine learning techniques have been utilised in various research in controlled environments to forecast and predict the growth of plants [22]. One of the primary advantages of machine learning techniques is their capacity to automatically solve non-linear tasks by utilising datasets from various sources. It offers a strong and adaptive framework for data-driven decision making that is widely used and has a lot of potential in agricultural applications [12]. In machine learning, regression is implemented using supervised learning method to predict continuous numerical values based on input features. In the plant yield prediction, yield is a continuous variable.
In a comprehensive survey that was carried out by Kwaghtyo and Eke [12] on smart farming prediction models for precision agriculture, they analysed different works that were carried out by different researchers on how machine learning models can be used in different areas of smart farming. In this survey, they found that both supervised machine learning which include support vector machine (SVM), decision tree, naïve bayes and unsupervised machine learning like ANN, convolutional neural network (CNN), k-means also ensemble methods have been helpful to automate smart farming practices. They also concluded that poor performance of some of the machine learning models was due to the nature of the data and the data preparations methods used. More details on SVR, LSTM and ANN are available in section 2.1.1, 2.1.2 and 2.1.3 respectively. This thesis included statistical analysis to ensure that the differences in algorithms performances were statistically significant and did not occur by chance. Cross validation is the technique that is applied in evaluating the performance of a machine learning by testing it on the unseen data. This is done by splitting the data into multiple folds, and the model is validated on a single fold while trained on the rest of the fold.

2.1.1 Support vector regression (SVR)

Support vector machine (SVM) is a frequently used algorithm for classification problems. In the classification problem, SVM will try to find out the optimal hyperplane which can separate the classes in the data with maximum boundary [28]. Support vector regression (SVR) which is a variant of SVM is designed by Vapnik and co-workers Bernhard Bose and Isabelle Guyon [17] to solve regression problems by predicting continuous or numerical output values based on the input features fed into the model. Similar to other regression algorithms, vector regression can also be applied for both linear and non-linear regressions. This is possible with the kernels in the support vector machine [17].

Support vectors are the vectors which represent the data points which lie within a margin called epsilon-tube from the hyperplane [28]. In training, SVR algorithm will try to find a function \( f(x) = wx + b \), which is as flat as possible with a maximum epsilon tube range(\( \epsilon \)) for the training of the data. The function \( f(x) \) represents the decision boundary of the SVR algorithm. The range of the epsilon-tube is determined by the support vector machine parameter epsilon(\( \epsilon \)) [28]. Inside an epsilon tube, the error for the data points is zero, but outside it penalises with corresponding values in the slack variables. During the training of SVR, the algorithm will try to minimise the objective function, \( \frac{1}{2} ||w||^2 + C \cdot \sum (\xi_i + \xi_i^*) \), here \( \xi_i, \xi_i^* \) are the slack variables, C is the regularisation parameter and w is the weight vector. The following are the important terms associated with support vector regression. Kernels are used in SVR to capture the relationship between the dependent and independent variables in the data [24].

**Kernels in SVR**

**Linear kernel:** linear kernel uses the kernel function \( K(x, y) = x \cdot y \) which is the dot product between the input features(x) and the target variable(y). This type of kernel is selected when the data can be linearly separable [28].

**Polynomial kernel:** Polynomial kernel function is \( K(x, y) = 1 + \sum (x \cdot y)^d \), where d is the degree parameter that determines the degree of the polynomial function. This kernel is applicable when the hyperplane has curved or non-linear shapes [28].

**Gaussian (RBF) kernel:** \( K(x, y) = \exp(-\gamma \cdot \sum (x - y)^2) \) is the kernel function of the radial basis function or gaussian kernel. Gamma, \( \gamma \) controls the kernel weight of the gaussian function. Gamma can be calculated by the equation, \( \gamma = 1 / (2\sigma)^2 \) [28] where sigma
is the standard deviation.

**Sigmoid kernel**: sigmoid kernel function is \( K(x, y) = \tanh(\gamma * x * y + r) \), here \( \gamma \) controls the curvature of the sigmoid function and ‘\( r \)’ is the bias of the tangent function (\( \tanh \)) [28].

### 2.1.2 Long Short-Time Memory (LSTM)

LSTM stands for Long Short-Time Memory which is a special type of Recurrent Neural Network (RNN) that can handle vanishing gradient problems. It was designed by Hochreiter and Schmidhuber [33] to resolve the problems in machine learning algorithms and traditional RNNs. LSTM is a network that is suitable for handling long term dependencies in sequential data such as time series, speech, and natural language data and alongside it has the capacity to store information for a longer period due to its memory cell. The architecture of a LSTM node is shown below.

![LSTM cell architecture](image)

The LSTM network consists of key components which are: Input layer, LSTM cells and Output layer. The input layer is responsible for receiving input data in the network at each time step, the output layer generates the final output of the network by receiving the final internal state of the LSTM cells as input. The LSTM cells are the main building blocks in the network that helps to select update, reset, and output information at each time. Each of the LSTM cell has three main components: Forget gate, Input gate and output gate. The forget gate decides and selects the information to be removed from the cell at each time step, the input gate controls and selects which new information to be added to the cell at each time step while the output gate passes the updated information from the current time step to the next time step.

The LSTM formula for a single time step can be expressed as follows:

- **Forget Gate**: \( f_t = \text{sigmoid}(W_f * [h_{(t-1)}, x_t] + b_f) \)
- **Input Gate**: \( i_t = \text{sigmoid}(W_i * [h_{(t-1)}, x_t] + b_i) \)
- **Candidate Memory**: \( \tilde{C}_t = \tanh(W_c * [h_{(t-1)}, x_t] + b_c) \)
- **Updated Memory Cell**: \( C_t = f_t * C_{(t-1)} + i_t * \tilde{C}_t \)
- **Output Gate**: \( O_t = \text{sigmoid}(W_o * [h_{(t-1)}, x_t] + b_o) \)
- **Hidden State**: \( h_t = O_t * \tanh(C_t) \)

where  \( x_t \) is the input at time \( t \), \( h_{(t-1)} \) is the hidden state from the previous time step, and \( h_t \) represents the hidden state of the current timestamp. Also, LSTM has a cell state that is represented by \( O_t \) as the output gate, \( h_t \) is the output, \( C_{(t-1)} \) and \( C_t \) for previous and current timestamps respectively, while the \( W \) and \( b \) terms represent the weight and bias parameters of the gates [39]. Some activation functions are being used by the LSTM which includes sigmoid, hyperbolic tangent functions that aids in the control of how information is passed across the
network and helps to prevent the problem of vanishing gradient that is common in traditional RNNs.

2.1.3 Artificial neural network (ANN)

It is an AI field and one of the important tools in machine learning. It is inspired by the function of a biological neural network that intends to mimic the human brain. ANN is a computational model that consists of neurons known as artificial neurons which are connected to one another and structured in various layers [7]. They are good tools in machine learning due to its capacity to handle complex tasks that are difficult to solve by humans teaching the machine to learn.

The architecture of Artificial Neural Network consists of three layers which are: Input, hidden and output layer. Input layer is the layer where input data is accepted, the hidden layer is between the input and output layers and it identifies patterns by performing mathematical operations on the input data [1]. The output layer conveys the output results of the transformed data in the hidden layer.

In the ANN block, the network receives input signal in vector form from the external source and gets multiplied by the corresponding weights to achieve the weighted sum, this sum passes as an input to the activation function to give the output.

\[ x_1.w_1+ x_2.w_2+ x_3.w_3+ \ldots \ldots \ldots \ldots \ldots x_n.w_n \]

Where: \(x_1, x_2, x_3..x_n\) represent the input data
\(w_1, w_2, w_3...w_n\) represent the weights

The activation function helps to introduce non-linearity to the network and convert an input signal in a Node to an output signal in an ANN network [7]. Thereafter, use this output as input in the layer that follows. The output of each neuron will be normalised by the activation function. There are different types of activation functions that are used in ANN which are categorised under Linear Activation Function or Non-Linear Activation Function and this activation is selected based on the task.

How the ANN works:

In the network, weights are initially assigned randomly to the input and are needed to give the output of the neural network. Information is fed into the network during the training whereby output value is achieved. Thereafter, the actual output and predicted output is compared, and the difference is calculated, which is known as error and this is used to update the weights. To minimise the error or remove the errors completely, the training set is fed into the network over and over until the error is drastically reduced or completely zero.

2.1.4 Performance metrics

Mean absolute error (MAE), mean squared error (MSE) and r-squared are the performance evaluation metrics that were commonly used for the regression models. MSE computes the squared difference between expected and actual values and computes the average of that difference and it is successful in picking the best models since it is particularly susceptible to outliers [8]. This is due to the MSE’s application of the L2 normalisation principle, which provides more weight to outliers [8]. In the training of machine learning models, it is very suitable as by minimising the MSE we can reduce the information loss [8]. The mathematical representation of MSE is given below.
MSE = \frac{1}{N} \sum (y - \hat{y})^2

where N is the number of data points in the test data. y, \hat{y} represents actual and predicted outputs respectively.

MAE is a measure of average errors between actual and predicted values. Here in this measurement, the absolute value of the difference is used which is very useful in interpreting the models results [8].

MAE = \frac{1}{N} \sum |y - \hat{y}|

R-squared is a fitness measure which shows how much of the variance in the data is learned by the model. In a regression model, it indicates the amount of variance in the dependent variable explained by the independent variables [2]. The value of r-squared ranges from 0 to 1. Here 0 means that the model fails to comprehend any variance in the dependent variable while 1 means the model is able to comprehend all the variance within the dependent variable.

r-squared = 1-(RSS/TSS)

where RSS is residual sum of squares and it can be calculated using \sum (y_i - \hat{y}_i)^2 [2]. Here the difference between each dependent with the corresponding predicted value is calculated. TSS is total sum of squares which can be calculated using the equation \sum (y - Y)^2 [2]. It is calculated by finding the sum of the squared difference between each dependent variable with the mean of the dependent variable (Y).

2.2 Research background

This section gives an outline about the previous studies conducted on plant yield prediction and the importance of tuning the hyper parameters. There have been many studies in predicting the yield of plants using machine learning algorithms. According to the study conducted by Suebsombut et al. [29] LSTM and Bi-directional LSTM were used to predict the soil moisture content to manage the resources. The data for this study was collected using sensors from the greenhouse in Chiang Mai province, Thailand. During the data preparation phase, mean and simple imputation were used to replace the missing values. After scaling the data, 14 features with 17,749 samples were used for the model implementation. For the proposed models, there were 4 LSTM and drop out layers, this is followed by output layers which consists of 12 dense units. Hyper parameters like learning rate, train test split ratio, batch size, time steps were tuned using empirical study. The average score of 5 fold cross validation was used to find the model performance. It was seen in the results that the LSTM model performed well in the prediction with an RMSE of 0.72 while Bi-LSTM has 0.76.

Farizan, Putrada, and Pahlevi [6] conducted a study to predict lettuce growth using SVR. The information is gathered from an experimental aeroponic device that has been set up in a controlled setting. Sensors took data for 21 days at a rate of once every minute. The data contains features such as air temperature, humidity, light intensity, water humidity, day, and light accumulation. Sensor data and observation data were combined for the investigation. Following cleaning, standard scalar was used for feature scaling in the processing phase, and the approach for feature selection was Pearson correlation. The features of water temperature, air temperature, humidity, light intensity, fresh weight, number of leaves, length of leaves, width of leaves, day, and light accumulative were chosen based on the value of the Pearson
correlation. Four SVR models were implemented to predict fresh weight, the number of leaves, leaf length, and leaf width, and five test scenarios were created and evaluated to get the best model for each target based on each feature’s score as determined by the Pearson correlation method. The first scenario included all features, and for the second, one feature was removed, followed by two, and so on for a total of five scenarios. The authors came to the conclusion that the accuracy and target are influenced by the features and target data that was selected. The SVR's accuracy for the number of leaves in scenario 1 had a maximum accuracy of 98.5% with all the features. The second scenario's best accuracy, 98.51%, is for the number of leaves. The leaf length in the third scenario had the highest accuracy, at 98.32%. The best accuracy is for leaf length in the fourth scenario, where it is 98.45%, and for leaf length in the fifth scenario, where it is 98.15%. The final performance of the four SVR models was evaluated with R-Squared and RMSE. The model for the leaf target had an R-Squared value of 0.99 and an RMSE of 0.15. The R-Squared value of the model for the fresh weight target is 0.97, while the RMSE value is 0.11. The R-Squared value of the model for the target leaf width is 0.98, and the RMSE value is 0.12. The R-Squared value of the model for the target leaf length is 0.98, and the RMSE value is 0.13. According to test results, the target features have optimum R-Squared values of 0.98 each and an average RMSE of 0.13 for the performance of the SVR model in predicting targets.

The authors of [3] carried out research on plant growth (plant height) prediction models for lettuce in plant factories using ANN. The data for this research was collected from a plant factory where nine Lettuce variety (Lactuca sativa) was grown. The data was collected for the whole plant growth cycle of the lettuce plant which is about 45 days. One growth cycle is considered as a sample, and for this research they used 36 samples for training and 9 samples for testing. The features used are air temperature, air relative-humidity, light intensity, electrical conductivity and plant height. The data was normalized before feeding it to the ANN model. Different numbers of nodes were used at the hidden layers which are 1 to 7 nodes. The evaluation result gave ANN with network architecture 4-7-1 as the best prediction model with RMSE of 0.032.

In research by Chaoya et al. [10] on Autumn crop yield prediction using data-driven approaches, explanatory factors and feature selection were done using a redundancy analysis (RDA). The interpretation rates of the explanatory factors were assessed using the basic effects of RDA. The conditional effect of RDA was explored to choose the features of the explanatory component. In this research, data from 2007 to 2016 was taken from the official website of National Bureau Statistics of China. The prediction models implemented are SVR, random forest regression (RFR) and deep neural network (DNN) to estimate autumn crop yields and these models were tested with coefficient of determination (R2), the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). It was seen in their result that the DNN gave the best performance among the three models when it was used on small data of 80 samples. It was also seen from their results that there is a good yield of the autumn crop, and this could be interpreted that the interpretation rate of the exploratory factors ranged from 54.3% to 85.0%.

Another study by Alhnaity et, al. [4] employed machine learning models like LSTM, SVR and random forest for predicting tomato yield and ficus benjamina stem growth. The data for this study was collected from two green houses in United Kingdom and Belgium for the period of 2017 to 2021. For the tomato yield prediction, climate data, manual yield measurement were used. Then the climate data was collected on hourly and weekly basis and the data was collected to a daily basis through data augmentation and implementation. The data was split into train, test and validation with 60, 50 and 25 percentage respectively. Three metrics MAE, RMSE and
MSE were used for the model evaluation. It was found that LSTM outperformed SVR and random forest with an MSE of 0.002, RMSE of 0.047 and MAE of 0.03 for tomato yield prediction. LSTM model gave the best performance in the ficus benjamina stem growth prediction with an MSE of 0.001, RMSE of 0.042 and MAE of 0.03. In this study, it was concluded that LSTM had a better performance than the machine learning algorithms.

Weerts, Mueller and Vanschoren [38] conducted an experimental study on the importance of hyperparameter tuning in machine learning algorithms. They chose 59 datasets from the OpenML-CC18 for the analysis. SVM (Support Vector Machine) with radial basis function and RF (Random Forest) are the algorithms employed in this experiment. The hyperparameters gamma, C, tol, shrinkage, and kernel were employed in the SVM, while in the random forest bootstrap, criteria, max_features, min_samples_leaf, min_samples_split, and n_estimators were chosen. The accuracy and overall averaged area under the ROC Curve (AUC) of the models were used to calculate their performance. Two experiments are part of their methodology. The authors employed a simple heuristic approach to discover the default parameter, and to avoid data leakage, they used a leave one out strategy in the default configuration of the parameters. They determine the hyperparameter setting that is repeated in these results from the selected 10 hyperparameter settings that have the best performance in this first experiment, and this could be set as the default value. In the second experiment, they apply nested cross validation with fixed and non-fixed conditions for the hyperparameters to determine the impact of tuning. The authors discovered that certain hyperparameters, such as the number of iterations in random forest, perform similarly to tweaking when left at their default values. While there are parameters with significant tuning risk that should be considered when implementing machine learning algorithms, SVM’s gamma and random forest’s max_features fall into this group.

Three machine learning models SVR, LSTM and ANN were selected in this thesis work. The combined dataset has various features that affect crop yield, and these algorithms are suitable in learning the non-linear relationship between the variables [28, 39, 32]. Furthermore, earlier studies on this particular subject with these machine learning algorithms demonstrated high accuracy in yield predictions, making them suitable for this experiment. One advantage of SVR is its ability to map data points to a higher dimensionality space [13]. Most algorithms are limited by the dimensionality of the input space, whereas SVR kernels can map data points to higher dimensionality. Furthermore, SVR’s kernels allow it to understand the underlying patterns in the data by selecting the proper kernel. The margin within the SVR also aids in minimizing the influence of outliers [13], making it appropriate for real-world data, such as the data utilized in this experiment. This made SVR one of the algorithms used in further exploration of RQ3. SVR have well defined parameters which have a direct impact in the model’s performance and the flexibility of the kernels makes it suitable for RQ1. In this thesis, the yield of lettuce is predicted for the days between a particular timeframe in the DAT stage, and the lettuce yield is influenced by the growth stage, this sequence repeats every 35 to 40 days. LSTM was appropriate for this situation due to its ability to manage sequential data and long-term interdependence within sequences. Even with little amounts of data, an LSTM is capable of accurately capturing temporal dependencies within prior values that could affect future results [39]. Most algorithms have difficulty with this, but the LSTM architecture was designed with this advantage [39]. ANN is popular for its ability to grasp complex patterns and handle non-linear relationships within the data. It is specifically useful when dealing with datasets with a large number of features, because ANN can automatically learn which features influences the prediction task, which will increase the model’s efficiency and accuracy [32]. The ANN’s scalability will also benefit future exploration of this subject with large data [32].
Thus, the ANN algorithm is a good fit for this thesis problem. It is also a good choice for exploring RQ2 due to its complex architecture and various hyperparameters that can influence model performance.

Furthermore, this thesis was influenced by the researches [31] and [38] in experimenting with different hyperparameter settings in the developed models. Tuning hyperparameters is critical in the implementation of machine learning models. Adjusting the hyperparameters rather than utilizing the default value will have an impact on the models' performance and generalization ability [38]. Tweaking the algorithm's hyperparameters is critical for finding the correct balance between the algorithm's complexity and performance [38]. While hyperparameter tuning, will also provide insight into how the models react under various parameter settings. Even while hyperparameter tweaking is crucial, it has a significant computational cost regarding the time taken to tune the models and the large amount of machine resources utilized. Hence in the parameter tuning of ANN and LSTM models, the number of nodes was chosen for hyperparameter tweaking experiments in this particular dataset. In the case of SVM, the hyperparameter chosen is the algorithm's kernel.

3 Method

The goal of this research is to predict the plant yield by applying machine learning algorithms and determining the best performance model. The state-of-the-art literature review methodology is used in this study to gain more knowledge in the previous work carried out by other researchers. Research materials for this project were found in a variety of databases, including Science Direct, Google Scholar, IEEE-Xplore, Research Gate, Scopus, and other databases, in order to obtain an understanding of the earlier studies that had been carried out on plant yield prediction. The research papers are found by searching in the categories of publications such as journals, open research articles, conference papers etc. Using plant yield prediction with machine learning, artificial intelligence in smart farming, deep learning in vertical farming, machine learning in hydroponics, support vector regression in yield prediction, artificial neural network as search terms.

Experiment is the systematic approach that was used for this thesis to answer the problem and the research question RQ3 in which we needed to implement, evaluate, and analyse and draw conclusion from the prediction of yield by the selected machine learning models. The benefit of this methodology is being able to have the capacity to explore model parameters since the researcher can make decision based on the experiment requirement and this was helpful to answer RQI and RQ2. In an experiment, the researcher has total control over how the experiment is carried out, which methodologies are used, and even the inclusion of extra procedures to make the experiment complete [30]. However, one disadvantage of an experiment is that its scope is limited [30]; the experiment is planned in such a way that it meets the requirements for the goal or the objectives of the research. In the case of interviews, they are employed when an in-depth qualitative insight is required [30]. Since this thesis is a deeper investigation to provide quantitative data that can give insight into how the yield can be increased, therefore experiment is suitable for this research rather than an interview. A case study is an alternative way of carrying out this research. A case study is an in-depth investigation into a single case at a specific period of time that can be used for comparative studies, because it is done on a specific example, the generalization of the results of case study is limited [30]. Moreover, case studies need more time and resources for a more in-depth analysis [30], which is not feasible in this research due to time and resources constraint.
The data science approach that was implemented for the steps required in solving this task is the Cross Industry Standard Process for Data Mining (CRISP-DM) method. CRISP-DM is a framework for planning and managing a data science project and it consists of six incremental stages that are iterable whereby the business problem can be properly answered before the completion of the whole cycle [9]. The first stage in CRISP-DM is the business understanding, the discussion with the key people in the farm explained about the data collection and the task they want to solve. The company's task was to develop a model that can predict the plant yield after the first 7 days in the DAT stage. The second stage of this CRISP-DM is understanding the data, and the discussion with the company representative gave insight to the features of the data provided. From the 4 datasets given, manual, climate and light recipe data were used for this experiment. Water data was left out because our experiment is carried out on the basis of growth cycle, and there are many missing values within the growth cycle that can only be replaced by linear interpolation, and this can cause inaccurate relationships within the data. The third stage is the data preparation which entails merging the datasets together, handling the missing data, cleaning the inconsistency in data to achieve a quality data that can be worked on. Then the records from day 8 to 22 in DAT stage was selected for the experiment. Thereafter, the dataset was split into test and train by using one sample of the DAT stage as the test data and the rest of the data was used for the training of the model. The fourth stage is the modelling, and three models were implemented which are SVR, LSTM, ANN to predict the plant yield. The poor performance of the models in this stage led to the exploration of 2 scenarios within this experiment. In the first scenario, records from day 8 to day 22 of the DAT stage with 9 features were selected (dataset 1) for the experiment. And in the second scenario, all the records of the DAT stage were used for the experiment (dataset 2). After this stage, was the evaluation stage and MSE, MAE and r-squared were used as the metrics to evaluate the implemented models. To respond to RQ1, hyper parameter adjustment using gridserachcv for different kernels in SVM was employed. The ANN and LSTM models were tested with different nodes to answer RQ2 at this level. RQ3 was answered using statistical analysis and performance scores. The final stage in the CRISP-DM is the deployment phase, and in this stage this research will be handed over to Ljusgårda AB for the solution to the task given.

4 Implementation of experiment

4.1 Dataset

This study aims to develop machine learning models to predict plant yield using the features in the data that was provided. Currently the company uses a manual estimation by the cultivation experts to predict the yield. The data for this research was obtained from the Ljusgårda company from their indoor farming facility in Tibro, Sweden and the data was collected both manually and through sensors that are attached to the growth pod. For this experiment, we selected the 'Examino' lettuce plant according to the company's request. The manual dataset (table 9 in appendix A) is collected manually by the workers in the company throughout the growth stage of the lettuce plant across all the growing rooms. The manually collected data starts from 2022-08-24 to 2023-03-23 and have 36103 records with 15 features. Based on the discussions with the company, ten features were removed since more than half of them had missing values, and the company advised against replacing those missing values with mean value because the data is unpredictable across the growth cycles. For better understanding of the features to the users, weight packable and day after transplant were
renamed as yield and growth cycles respectively. The selected five features from the manual datasets are timestamp, grow-room, growth-stage, yield and weight_waste. The data was grouped on daily basis to avoid inconsistencies in the datasets.

The manual data has a dimension of (505,5) after the data cleaning. The climate information (table 1 in appendix A) for each grow room was in 15 csv files, each with six features and these files were merged to create climate dataset. This dataset consists of information of both humidity and relative humidity and the relative humidity was chosen based on inputs from the company. Climate datasets were also organized daily, and the features selected are timestamp, grow room, temperature, CO2 and relative humidity and this data was merged with manual dataset. 3225 records and 5 features were left after cleaning the climate datasets. The light recipes (table 11 in appendix A) had 14 features, and 5 important features with three light recipes were selected after discussion with the company. The selected features of this light recipe are blue, red, green, far-red and total intensity and these were measured in μmol/s/m². Thereafter, manual, climate and light recipe data were merged together to form a single dataset consisting of 504 records and 10 features. After the exploratory data analysis (EDA) (Fig 1 in Appendix A), blue and total intensity were dropped from the features of the dataset to avoid the multi collinearity. For this experiment, we considered the growth cycle from day 1 to day 22 since there was only one record for the yield on day 23. To explore the scenarios mentioned in section 3, the dataset was split into two datasets. For the first scenario, the combined dataset was filtered so that the data has records from day 8 to 22 of DAT cycle only and this is named as dataset 1. The whole dataset consisting of the full growth cycle from day 1 to 22 of DAT cycle was used to explore the scenario 2 and named as dataset 2. Thereafter, the datasets were used for model training and evaluation. Most of the related studies [10], [4] have more than three years of data and this contributed to the better performance of their models. Throughout our experiment, we found that our models have a poor performance when compared to these previous studies because we only have data of about seven months. The selected features for this experiment are listed below (Table 1).

<table>
<thead>
<tr>
<th>Dataset features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>growth_stage</td>
<td>Growth stages of the plant in the form of day after sowing (between 1 to 23)</td>
</tr>
<tr>
<td>Yield</td>
<td>Plant weight at each of the growth stage in grams</td>
</tr>
<tr>
<td>temperature</td>
<td>Temperature maintained in the room in degree celsius.</td>
</tr>
<tr>
<td>CO2</td>
<td>CO2 in the room for that particular day</td>
</tr>
<tr>
<td>Relative humidity</td>
<td>Humidity in the room for that particular day</td>
</tr>
<tr>
<td>green</td>
<td>The intensity of green light in μmol/s/m².</td>
</tr>
<tr>
<td>red</td>
<td>The intensity of red light in μmol/s/m².</td>
</tr>
<tr>
<td>far-red</td>
<td>The intensity of far-red light in μmol/s/m².</td>
</tr>
<tr>
<td>grow-room</td>
<td>Grow room number ranges from 1 to 15</td>
</tr>
<tr>
<td>Weight_waste</td>
<td>Wastage of the plant during harvesting in grams</td>
</tr>
</tbody>
</table>

Table 1. Features in dataset

4.2 Implementation of SVR (RQ1)

This SVR model was implemented using jupyter notebook version 6.3.0 and sklearn version 1.2.0 with the approach in 2.1.1. Gridsearchcv module from sklearn library was used to
implement SVR models with different kernels. For all the models that were implemented, two versions of models were built using dataset 1 and dataset 2 respectively to predict the plant yield. Gridsearchcv algorithm will exhaustively trying out all possible combinations of the given parameters within the model to find a model with optimised result. Gridsearchcv in the sklearn library was used to tune the hyperparameters of SVM models with different kernels. The number of cv folds in the method was gradually increased to improve model performance. The cross-validation folds of the gridsearchcv were set to 15 in the final implementation for all SVR models. This value was discovered by gradually increasing the number of CV folds from a smaller number. To answer RQ1, different kernels of SVR were experimented to find the effect of kernels in the performance of SVR model. The implemented SVR models design is given in table 2.

<table>
<thead>
<tr>
<th>SVR</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kernel</td>
<td>RBF kernel</td>
<td>RBF kernel</td>
</tr>
<tr>
<td>C</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>epsilon</td>
<td>0.2</td>
<td>0.01</td>
</tr>
<tr>
<td>gamma</td>
<td>0.05</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 2. SVR model design for dataset 1 and 2

4.3 Implementation of LSTM (RQ2)

This model was also implemented using the same jupyter notebook version with TensorFlow version 2.7.0 and keras version 2.7.0. Two LSTM models were implemented with the approach mentioned in 2.1.2 using dataset 1 and 2 respectively. The models have three layers: one input layer, one hidden layer and one output layer. For dataset 1, the input layer has 15 LSTM units to handle growth cycle from day 8 to day 22 which is 15 days. While the datasets 2 have 22 LSTM units in its input layer to train the whole growth cycle from day 1 to day 22 which is 22 days. In addition, the number of epochs, nodes in the hidden layer and the batch size were optimised using the gridsearchcv. Since the tuning of the neural network is difficult in computation time, the cv folds of both LSTM and ANN in gridsearchcv was set to 5 folds and the number of folds fixed to 5 after gradually increasing the cv folds in gridsearchcv. Then model was experimented by gradually increasing the number of nodes in the hidden layer to study how the model behaves with change in the number of nodes, and this answers RQ2. In this parameter tweaking experiment of LSTM, epochs value and batch size value are fixed to 75, 16 respectively. The implemented LSTM models design is given in table 3.

<table>
<thead>
<tr>
<th>LSTM</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of layers</td>
<td>Three layers: one input, one hidden and one output</td>
<td>Three layers: one input, one hidden and one output</td>
</tr>
<tr>
<td>Nodes</td>
<td>(15, 20,1)</td>
<td>(22,4,1)</td>
</tr>
<tr>
<td>Optimiser</td>
<td>Adam</td>
<td>Adam</td>
</tr>
<tr>
<td>Epochs</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>Batch size</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 3. LSTM model design for dataset 1 and 2
4.4 Implementation of ANN (RQ2)

Similar to the LSTM models, two versions of this model were also implemented with the approach mentioned in 2.1.3 using the same Jupyter Notebook, Tensorflow and Keras version. The ANN models were implemented with three layers, input, hidden and output using dataset 1 and dataset 2 respectively. The input layer was designed with 9 nodes since the number of input features is nine. The parameters of these models were fine-tuned using gridsearchcv. The implemented ANN models design is given in table 4. The models were also experimented with different number of nodes the same way as mentioned in 4.3 to study how the model behaves with change in the number of nodes and answers RQ2. Similar to LSTM in the parameter tweaking experiment the epochs and batch size were fixed to 75, 16 respectively.

<table>
<thead>
<tr>
<th>ANN</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of layers</td>
<td>Three layers: one input, one hidden and one output</td>
<td>Three layers: one input, one hidden and one output</td>
</tr>
<tr>
<td>Nodes</td>
<td>(9,20,1)</td>
<td>(9,16,1)</td>
</tr>
<tr>
<td>Optimiser</td>
<td>Adam</td>
<td>Adam</td>
</tr>
<tr>
<td>Epochs</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>Batch size</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 4. ANN model design for dataset 1 and 2.

4.5 Statistical Analysis (RQ3)

Cross validation was used for the statistical analysis and was implemented using Kfold class in sklearn. In this experiment, the data was partitioned into 5 folds for the cross validation. The number of folds in this experiment was set to 5 since cross validation is performed on all models, including the neural network, which requires extra computing time. Based on earlier research, we discovered that at 5 cross validation folds, the models perform well and were within the computational cost. Mean and variance of the MSE of each model were calculated across the cross-validation score of each model to get an insight into the model performance.

5 Results of experiment

The results in section 5.1 are for RQ1 regarding the influence of different kernels in the performance of SVR. Section 5.2, 5.3 results are for RQ2 based on different number of nodes that were explored. The experiment results in section 4 are related to RQ3 which is about statistical analysis.

5.1 Results of support vector regression (RQ1)

As mentioned in section 4.2, two datasets dataset 1 and dataset 2 were experimented with different SVR kernels. MSE, MAE and r-squared were the evaluation metrics used and the results of these experimentation are shown in table 5.
## 5.2 Results of long short-time memory (RQ2)

Two models of LSTM were implemented with two different dataset named dataset 1 and dataset 2 as mentioned in section 4.3. LSTM with a single hidden layer was implemented and tested with different nodes and evaluation metrics employed are MSE, MAE, and r-squared which are shown in table 6.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset 1(8-22 days of DAT)</th>
<th>Dataset 2(1-22 days of DAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation metrics</td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Linear kernel</td>
<td>153.12</td>
<td>9.35</td>
</tr>
<tr>
<td>RBF kernel</td>
<td><strong>134.12</strong></td>
<td><strong>8.9</strong></td>
</tr>
<tr>
<td>Sigmoid kernel</td>
<td>23.46</td>
<td>11.56</td>
</tr>
<tr>
<td>Polynomial kernel</td>
<td>153.50</td>
<td>9.36</td>
</tr>
</tbody>
</table>

Table 5. Result of SVR performance with different kernels

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset 1(8-22 days of DAT)</th>
<th>Dataset 2(1-22 days of DAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation metrics</td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td>4 nodes</td>
<td>381.53</td>
<td>15.58</td>
</tr>
<tr>
<td>8 nodes</td>
<td>375.71</td>
<td>14.87</td>
</tr>
<tr>
<td>12 nodes</td>
<td>357.31</td>
<td>14.87</td>
</tr>
<tr>
<td>16 nodes</td>
<td>335.47</td>
<td>14.63</td>
</tr>
<tr>
<td>20 nodes</td>
<td><strong>216.76</strong></td>
<td><strong>14.52</strong></td>
</tr>
</tbody>
</table>

Table 6. Experiment result of LSTM with different nodes.

## 5.3 Results of artificial neural network (RQ2)

This section contains the result of the experiment carried out on ANN as mentioned in section 4.4. The evaluation metrics used were MSE, MAE and r-squared, the results are shown in table 7.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset 1(8-22 days of DAT)</th>
<th>Dataset 2(1-22 days of DAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation metrics</td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td>4 nodes</td>
<td>183.64</td>
<td>10.89</td>
</tr>
<tr>
<td>8 nodes</td>
<td>153.50</td>
<td>9.30</td>
</tr>
<tr>
<td>12 nodes</td>
<td>124.54</td>
<td>9.03</td>
</tr>
<tr>
<td>16 nodes</td>
<td>121.35</td>
<td>9.01</td>
</tr>
<tr>
<td>20 nodes</td>
<td><strong>119.18</strong></td>
<td><strong>8.76</strong></td>
</tr>
</tbody>
</table>

Table 7. Result of ANN with different nodes
5.4 Results of statistical analysis (RQ3)

The statistical analysis results for the implemented models for both dataset 1 and 2 is shown in table 8. In the table the standard variation and mean of the mean squared errors of each of the model after the cross-validation results is shown.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Dataset 1</th>
<th>Dataset 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Std</td>
<td>95%confidence interval</td>
</tr>
<tr>
<td>SVR</td>
<td>63.79</td>
<td>127.58</td>
</tr>
<tr>
<td>ANN</td>
<td>12.90</td>
<td>25.8</td>
</tr>
<tr>
<td>LSTM</td>
<td>23.83</td>
<td>47.66</td>
</tr>
</tbody>
</table>

Table 8. Results of statistical analysis

5.5 Analysis of the joint results

Scatter plots in Fig 3 and Fig 4 shows the relationship between the predicted value and actual value of different models, SVR, LSTM and ANN for dataset 1 and 2.

Fig 3. Actual v/s predicted yield with dataset 1

Fig 4. Actual v/s predicted yield with dataset 2.

It was found from the Fig 3 and 4, for both datasets, that the implemented models gave a better
prediction when the yield is below 40. It could be because of high number of records that were present below the yield range of 40. On further exploration of the datasets, it was observed that for the yield to reach 40, the plant should at least reach day 15 in DAT stage. Between day 1 to 15, there were 347 records available in the dataset 2, which is almost 70% of this data. In the case of dataset 1, 221 records were present between day 8 to 15 of the growth stage, which is almost 60% of dataset 1. The models are giving better prediction below the yield range because the models have more data available for training below this yield range.

Multiple bar chart in Fig 5 and 6 shows the performance of SVR, LSTM and ANN for both dataset 1 and dataset 2. Each multiple bar graph in Fig 5 represents MSE, MAE, r-squared. The blue coloured bars represent dataset 1 and orange colour represents dataset 2.

![Fig 5. Performance of fine-tuned models with MSE, MAE, r-squared](image)

It was seen from the Fig 5 for dataset 1, the MSE of SVR and ANN ranges between 100 and 120 while LSTM have a high error for both datasets. In the multiple graphs for MSE, ANN has the least MSE error in dataset 1 and 2, while SVR has the least error in dataset 2. The plot for MAE showed that all the models have less error when its working with dataset 2. All the models with dataset 2 have high r-squared which shows the high fitness of the models when they are working with dataset. LSTM gave a high error in terms of MSE and MAE for both datasets.

Fig 6 shows the comparison of the MAE between the models with confidence interval. In the case of dataset 1 ANN had the least MAE and also have the least confidence interval. For dataset 1, MAE of SVR is almost same as ANN but with a little difference of 0.14, while the confidence interval of MAE for SVR is higher than that of ANN, and this makes ANN suitable for dataset 1. For dataset 2, SVR has the least MAE and all the implemented models have almost similar confidence interval of MAE. This made SVR suitable for dataset 2.
In dataset 1 which has day 8 to 22 of the growth cycle, ANN gave the least error with an MSE of 111.56 and MAE of 8.34 compared to the other two models implemented. In dataset 2 which has day 1 to 22 of the growth cycle, SVR has the least error with an MSE and MAE of 111.95 and 6.11 respectively. We found that with dataset 1 ANN model performed better, while in dataset 2, SVR performed better. Statistical analysis was done to verify these findings, and the result is shown in Fig 9 and Fig 10.

Fig 7 shows multiple bar charts of the performance of different kernels in SVR. Similar to Fig 5, blue colour represents dataset 1 and orange colour represent dataset 2. The figure consists of three multiple bar charts, and each represent MSE, MAE and r-squared respectively.

Fig 7. Performance of SVR with different kernels

In the above figure, it was seen that all the SVR kernels implemented with dataset 2 have a better performance than dataset 1. The sigmoid kernel has the highest error in both datasets. In dataset 1, RBF kernel have less error compared to other kernels which is followed by the linear and polynomial and sigmoid kernels. In the case of dataset 2, linear kernel has the least MSE, while the least MAE is RBF kernel.

Fig 8 shows the results of experiments with different nodes in LSTM and ANN models. ANN and LSTM with dataset 1 is represented by blue and green colours respectively while ANN and LSTM with dataset 2 is represented by yellow and red colours respectively.

Fig 8. Performance of neural network models with single hidden layer

In Fig 8, it was seen that in dataset 1, the increase in number of nodes in the hidden layer, the error is decreasing. In dataset 2, LSTM increases with increase in the number of nodes and in ANN, there is smaller effect when the nodes were increased.

Fig 9 and Fig 10 shows the result of statistical analyses of the models implemented on dataset 1 and dataset 2 respectively. The y axis represents the average MSE of 5-fold cross-validation.
The error line represents the 95% confidence interval of the MSE scores.

![Fig 9. statistical analysis of models implemented with dataset 1](image1)

Fig 9. statistical analysis of models implemented with dataset 1

![Fig 10. statistical analysis of models implemented with dataset 2](image2)

Fig 10. statistical analysis of models implemented with dataset 2

Out of the models implemented with dataset 1, ANN had the least MSE across the 5-fold cross-validation and a smaller range of uncertainty. In dataset 2, SVR has the least MSE but with higher uncertainty.

6 Discussion

6.1 Comparison to Previous Research

Most previous studies used various machine learning algorithms, including those used in this study, the most widely used method being LSTM or its variants. The majority of these researches were conducted using climate data or plant dimensions such as plant height, leaf area, and so on. However, for our experiment, the light recipe combined with climatic information was used, and in the case of the models implemented in this research, hyperparameter tweaking experiment was employed to study the behaviour of models for this type of data. Most previous studies had at least three to five years of data, and LSTM was the one that performed well in these studies when it was employed as one of the many models. However, in our study, we had only a few months of data, and the model that performed poorly across all scenarios was LSTM. Furthermore, we were able to discover the effect of hyperparameters in the selected models for this particular problem through the experiment, which was a previously unexplored area. The implemented machine learning models, as well as the documented hyperparameter tweaking process and results, were contributed to domain experts and researchers for having a deeper understanding of model behaviour (the right balance between model complexity and interpretability) facilitating model selection, interpretation, and optimization. In the agricultural industry, the capacity for prediction of the
model can help provide practical suggestions that enable farmers to improve productivity while avoiding resource wastage.

6.2 Methods, implementation, and results

Based on the data provided by the company, we conducted a review on previous studies in this area. The most challenging part was model selection that is suitable for the problem proposed by the company. It was discovered that employing a single model to address this problem would be challenging, owing to the huge variety of machine learning algorithms that could be employed for regression. During the EDA, it was seen that there was non-linearity within the data (Fig 12 in appendix A). While a simple linear model can provide an easy-to-understand baseline, the complexity of machine learning models enables them to learn the complicated relationships within the data. Even with small data, machine learning models have the ability to learn the non-linear relationships within the data, with parameter tweaking these models can be shaped to a best fit for the data \[12\]. Machine learning algorithms are better choice than a simple linear regression model, thus machine learning models were selected. Further research about the benefits (Sections 4.2, 4.3 and 4.4) that different models can bring, we chose the machine learning models namely SVR and LSTM. After confirming the features that can be employed, we proceeded with the selected models as a possible solution to the research problem. Furthermore, researching on the model behaviour using hyperparameter tweaking was identified as another objective of this thesis project.

In order to obtain accurate predictions over the time period of day 8 to 22, the data was aggregated on a daily basis, which aids the model in recognizing daily variations or changes in the feature that may affect the yield. Following the discussion with the domain experts, the features that was discovered to be multicollinear were identified and removed as described in section 4.1. In the initial implementation stage, the algorithms LSTM and SVM did not perform well. Thus, after talking with the domain expert, other suitable methods were looked into and it was decided that the whole growth stage for the model training can be investigated. Furthermore, an additional algorithm ANN was added into this experiment because it can discover non-linear correlations between the variables \[11\]. For this experiment, two scenarios were considered; data from days 8 to 22 (dataset 1) and the entire growth stage from days 1 to 22 (dataset 2). For model training, all chosen features were applied to both scenarios. In selection of suitable evaluation metrics for these models, it was intended to use MSE to identify the best model because it is highly susceptible to outliers and imposes a large penalty for greater error. With multiple models, MSE is useful in identifying the best performing one but when it comes to interpreting the prediction results, it is not suitable enough since deviation of the prediction from the actual value can only be realized through the square root of the MSE. Since MSE gives high penalty for the higher error, the square root MSE is not suitable for showing how much our model’s prediction is far from the actual value \[8\]. Then after discussing with the domain expert from the company, it was decided to include MAE as it has a high interpretability of the results since it is calculated in the same units of the original dataset. When it comes to explaining the performance of the models, the company also prefer MAE as it is less sensitive to the outlier which measures the average error of predictions, makes it suitable in explaining how the outcome of the model varies from the original data.

Following the implementation of the three models SVR, LSTM, and ANN for datasets 1 and 2, it was discovered that the models performed better with dataset 2 because it comprises the entire growth cycle. It provides the models with more data to work with that includes the early
days of the growth cycle which have a strong influence in yield prediction. Therefore, it is important to consider the whole growth cycle of the plant in yield prediction. ANN had the lowest MSE of all the models employed. However, SVR's error was also quite close to ANN's and it has the lowest MAE with dataset 2. The r-squared of the implemented models for dataset 2 was greater than for dataset 1, indicating that giving the models more data offers them a better chance of learning the variation within the data points. After performing statistical analysis, it was discovered that ANN had significantly less uncertainty than SVR, making it appropriate for this thesis problem which is lettuce yield prediction. One further intriguing observation was that, even while LSTM had the highest error with dataset 2 of all the deployed models, it had the lowest range uncertainty. This may be a result of the LSTM's ability to learn the temporal dependencies present in the data, which allows it to learn the sequential patterns present in the data even though the individual predictions are far apart. After speaking with the company's domain experts and reviewing the results of the experiments, it became clear that the MAE of ANN and SVR, which are 6.23 and 6.11 respectively, is very favourable than company's manual yield estimation by the cultivation experts. According to domain experts, the ANN and the SVR with rbf kernel are suitable for yield prediction since they provide results that are nearly comparable in terms of MAE and the models are better solution than that they currently have. They also expressed that it can only be implemented after the models are verified with the new data.

Another objective of this thesis work was exploring the hyperparameter tweaking. In the case of SVR, the parameter that was considered was the kernels, the experimentation results showed that the kernel with the most complex function is RBF and is able to learn the non-linear relationships between the data points. The performance of the kernels is influenced by the kernel complexity. Out of all the kernels implemented, RBF kernel was the kernel with most complex function and with least MAE error, it was true for both the datasets 1 and 2. This is followed by the linear and polynomial kernel with degree one which are the simplest kernels implemented. The least performed kernel is the sigmoid kernel which is the most difficult to implement as it relates on sigmoid coefficients and the kernel width. The number of nodes in hidden layer was explored with ANN and LSTM in this thesis and had an impact on how the neural networks performed with datasets 1 and 2. With dataset 1, ANN's performance improved as the number of nodes increased and achieved stability at 12 nodes. However, with dataset 2, adding nodes did not significantly improve the ANN's performance, and after 16 nodes, the model's error rate started to increase. ANN typically demonstrates that as the number of nodes increased, the model's performance gets improved, but after a certain number of nodes, the performance became worse. The LSTM model had a different performance in dataset 1 and 2. Compared to dataset 1, which got worse as the number of nodes increased, dataset 2 got better as the number of nodes increased. This demonstrates the significance of performing hyperparameter tweaking in identifying the best hyperparameter configuration for each dataset.

The main limitation of this thesis work was the small data provided by the company. Based on previous studies, at least three years of data were required to accurately predict plant yield; however, we only had seven month’s worth of data, which was insufficient for developing a suitable model that could solve the company’s task. In addition to the limitation, the dataset provided had varying numbers of records for the growth stages. The first 15 days of the DAT stage had more entries in the data, which resulted in high accuracy for predicted values in this range. In contrast, the later stages had more inaccurate predictions, while the late stages prediction was crucial for improving the yield of the lettuce plant. The time constraint for this thesis was another limitation. This work required extensive exploration with different machine
learning models and parameters because it is an industrial level challenge but due to the time limit, we have only used three machine learning models for this challenge. Another limitation in this thesis is dimensional properties of plants data was not used due to the decision of the domain experts. The data contained features; canopy area, roots temperature, roots electricity, number of leaves and the reason for exclusion of these features was the presence of high number of missing values within the features. According to the domain expert, replacing the values that were missing in these features could result in the machine learning algorithm to learn the incorrect relationships between the features. The inclusion of the plant dimensions would have made the study more detailed and improve the accuracy of the results because they always play a significant impact in enhancing plant production.

Regardless of the limitations, the experiment’s results can give important insights into lettuce plant prediction, where series of predictions made at different stages of growth are taken into account. The results will also be useful for understanding the role that hyperparameter tuning contributes to ML algorithms performance.

6.3 Ethical and societal aspects

Ethical concerns and data privacy impact were strictly considered in the prediction of crop yield during this study. Crop yield data contains sensitive business information that competitors within the agricultural field and other unauthorized parties should not be able to access and ensuring confidentiality is very crucial in this aspect. Ensuring the privacy and confidentiality of the data collected and business information; consent for the use of data was obtained from the Ljusgårda company as part of the agreement process and access was limited to those individual working on this project. To further reduce potential privacy risk, minimum of data that was required for the prediction of crop yield was collected and permission were received to publish this research with the images provided to us, while the data and code will not be published.

7 Conclusion

The objective of this thesis work was to discover the solutions to the research objectives of how the selection of the kernel and the number of nodes in the hidden layer affect the performance of SVR, LSTM, and ANN, respectively, and finding the best performing model for the yield prediction. Three algorithms SVR, LSTM, and ANN were successfully employed, and parameter tweaking experiments were conducted. The experiment explored two scenarios, one with the growth stage from day 8 to day 22 (dataset 1), and the other with the entire growth cycle (dataset 2). The findings revealed that the full growth cycle significantly improved the model’s prediction. After experimenting with different kernels in SVR it was discovered that SVR with rbf kernel performed better, and it was found that the kernel complexity helps the models improve their performance. In the case of LSTM and ANN neural networks, it was observed that optimizing the number of nodes in accordance with the input data might improve model performance. While the LSTM performed differently for datasets 1 and 2, the performance of ANN improved as the number of nodes rose for both datasets. This illustrates that model performance can vary depending on a variety of factors, and hyperparameter tweaking can enhance the model’s fit according to the given data. SVR with a rbf kernel and ANN with 20 nodes in the hidden layer had the lowest MAE and MSE respectively among the implemented models, making them the best-performing algorithms.
In conclusion, the implementation of two models with high performance to solve the business challenge of predicting the lettuce yield, we have successfully provided answers to the research objectives. The three models' parameter tweaking experiments provide insightful information for enhancing model complexity and optimizing model suitability for the data.

7.1 Future work

There are several approaches that can improve this study which can be explored in the future. A detailed analysis of the factors impacting the prediction could have been done with more data. The implementation of a feature scoring approach to identify the key features and the gradual incorporation of the features to assess the variation in the model's performance can give good understanding of the impact of the features on the model performance. This would have been useful in the case of improving the plant's yield by identifying the influencing factors.

Another area that can be explored is including a more thorough investigation of the parameter tweaking in this study. Due to the time constraint, the parameter taken into consideration for this study was the kernel in the case of SVR and the number of nodes in the hidden layer in the case of neural networks. The performance of SVR is also influenced by parameters C, gamma, and epsilon. The number of hidden layers, learning rate, dropout rate, batch size, and numerous other hyperparameters can affect the performance of neural networks. Therefore, integrating these in this research will increase understanding of how parameters affect the behaviour of the models.

Developing an application that will help the data experts to understand the effects of altering the data of the independent features to the plant yield and also make it easier and less time-consuming. This extension can be done in the deployment stage of this work. Trained machine learning models can be stored and used in the backend of these applications, and the use of cloud technology would make it simpler to integrate and analyse the huge volume of data in these kinds of applications.
References


Appendices

Appendix A

Data and EDA results

<table>
<thead>
<tr>
<th>Dataset features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day after transplant</td>
<td>Growth stages of the plant in the form of day after sowing (between 1 to 23)</td>
</tr>
<tr>
<td>Air temperature</td>
<td>Air temperature manually measured close to plant in growing room in degrees</td>
</tr>
<tr>
<td>Canopy area</td>
<td>Area of the plants canopy</td>
</tr>
<tr>
<td>Roots_EC</td>
<td>The plants root’s electro conductivity measure</td>
</tr>
<tr>
<td>Roots_Temperature</td>
<td>The plants root’s temperature measured in degrees</td>
</tr>
<tr>
<td>Weight_packable</td>
<td>Plant weight at each growth stage in grams</td>
</tr>
<tr>
<td>Plant_Height</td>
<td>Plant height at each growth stage</td>
</tr>
<tr>
<td>Leaf_Count</td>
<td>Number of leaf count on each stage</td>
</tr>
<tr>
<td>grow-room</td>
<td>Grow room number ranges from 1 to 15</td>
</tr>
<tr>
<td>Weight_waste</td>
<td>Wastage of the plant during harvesting in grams</td>
</tr>
<tr>
<td>Batch id</td>
<td>Batch id represents the batch to which the plant belongs</td>
</tr>
<tr>
<td>Weight</td>
<td>Total weight of the plant before harvesting</td>
</tr>
<tr>
<td>leaf_temperature</td>
<td>Temperature below the leaf surface</td>
</tr>
<tr>
<td>humidity</td>
<td>`humidity maintained in the grow rooms</td>
</tr>
</tbody>
</table>

Table 9. Features of Manual dataset

<table>
<thead>
<tr>
<th>Dataset features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>grow-room</td>
<td>Grow room number ranges from 1 to 15</td>
</tr>
<tr>
<td>temperature</td>
<td>temperature manually maintained in growing room in degrees</td>
</tr>
<tr>
<td>Co2</td>
<td>Co2 measured close to the plant surface</td>
</tr>
<tr>
<td>Relative humidity</td>
<td>Amount of water in air in percentage</td>
</tr>
<tr>
<td>humidity</td>
<td>`humidity maintained in the grow rooms</td>
</tr>
</tbody>
</table>

Table 10. Features of Climate dataset

<table>
<thead>
<tr>
<th>Dataset features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light type</td>
<td>The company have different light types which is according to the stage most</td>
</tr>
<tr>
<td></td>
<td>commonly used is the standard light</td>
</tr>
<tr>
<td>Light recipe</td>
<td>Name of the light recipe used</td>
</tr>
<tr>
<td>blue</td>
<td>The intensity of blue light is used in the recipe in μmol</td>
</tr>
<tr>
<td>Feature</td>
<td>Description</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Green</td>
<td>The intensity of green light is used in the recipe in μmol</td>
</tr>
<tr>
<td>Red</td>
<td>The intensity of red light is used in the recipe in μmol</td>
</tr>
<tr>
<td>Far-red</td>
<td>The intensity of far-red light is used in the recipe in μmol</td>
</tr>
<tr>
<td>Total intensity</td>
<td>The total intensity of the light applied to the plant in μmol/s/m²</td>
</tr>
<tr>
<td>Grow room number</td>
<td>Grow room number where the recipe is applied</td>
</tr>
<tr>
<td>Day after transplant</td>
<td>Growth stages of the plant in the form of day after sowing (between 1 to 23)</td>
</tr>
<tr>
<td>Light on</td>
<td>The time at which the light turned on after light off</td>
</tr>
<tr>
<td>Light off</td>
<td>Time at which all the lights were turned off</td>
</tr>
<tr>
<td>Fade in fade out</td>
<td>The duration at which the light will be turned off and turned on again</td>
</tr>
<tr>
<td>End of applying</td>
<td>Date at which the recipe was stopped</td>
</tr>
</tbody>
</table>

Table 11. Features of light recipe dataset

Fig 11. Correlation between features in the dataset

Fig 12. Scatter plot between the features and yield
Appendix B

Contributions of individual members

1. **Data collection and data preprocessing:** This task is completed by both Anjali and Mary. The collected data was combined to for adjusted to make it suitable for the project.

2. **Research Questions:**
   - **Anjali:** How does the choice of the kernel in Support Vector Regression affect the performance of the model in predicting the yield?
   - **Mary:** In the neural networks -LSTM and ANN, how does the number of neurons in hidden layer affect the model's performance?
   - **Both:** Based on the data, which one of the implemented models has a better performance?

3. **Models and statistical analysis:**
   - **Anjali:** Implementation and hyperparameter tweaking of Support Vector Regression
   - **Mary:** Implementation and hyperparameter tweaking of LSTM and ANN
   - **Both:** Statistical Analysis

4. **Evaluation Metrics:** Three evaluation metrics MSE, MAE, r-squared were identified and used to evaluate the model's performances.

5. **Thesis writing**
   - Section 1 and 1.1 handled by Mary and Anjali
   - Section 1.2, RQ 1 handled by Anjali, RQ 2 handled by Mary RQ3 by both Mary and Anjali
   - Section 2, 2.1, 2.1.4, 2.2 handled by Mary and Anjali
   - Section 2.1.1 handled by Anjali
   - Section 2.1.2, 2.1.3 handled by Mary
   - Section 3 handled by Mary and Anjali
   - Section 4,4.1, 4.5 handled by Mary and Anjali
   - Section 4.2 handled by Anjali
   - Section 4.3, 4.4 handled by Mary
   - Section 5.1 handled by Anjali
   - Section 5.2, 5.3 handled by Mary
   - Section5, 5.4 and 5.5 handled by Anjali and Mary
   - Section 6,7, References, Appendices and Abstract handled by Anjali and Mary